

INTEGRATION OF CORRELATION BASED FEATURE SELECTION AND DECISION TREE CLASSIFIER TO IMPROVE THE THYROID DISEASE DIAGNOSIS

Nagavali Saka¹ and S. Murali Krishna²

¹ Research Scholar, JNTUA, Anantapur, Andhra Pradesh, India.

Email: vali214@gmail.com

² Professor, Department of Computer Science and Engineering, Sri Venkateswara College
of Engineering (SVCE), Tirupati, Andhra Pradesh, India.

Email: muralikrishna.s@svcolleges.edu.in

ABSTRACT

Autoimmune disorders are aexpansivearray of correlated diseases in which unsuitable immune reactions of the body risebeside its own organs, tissues and cells, ensuing in tenderness and mutilation. This response may disturb a specific tissue,organ or the complete body. One among the most common disorders isHashimoto's Thyroiditis, an autoimmune ailment in which the thyroid gland is slowlydamaged. According to statistics of Indian Thyroid Society (ITS), thyroid disorders are on the escalation in India. About 1 in 10 Indian adults agonize from hypothyroidism, a situation in which the thyroid gland doesnot yieldadequate thyroid hormones to meet the necessities of the body. This disorder is twofold as predominant in women as in men and is common amongst women of child-bearing age. The machine learning plays a crucial role in the practice of ailment prediction. This paper handles Correlation based Feature Selection and application of multi-class classification methods: Support Vector Classifier (SVC), k-Nearest Neighbour (k-NN) and Decision Tree (DT)in the diagnosis of thyroid disease centred on the statisticscollected from the dataset taken from UCI machine learning pool.

Keywords—Thyroid, Prediction, Pearson, Correlation, Machine Learning, Classification

INTRODUCTION

Autoimmune Thyroid Disease is ever morebeheld as a continuum not only comprising distinct ailment entitiesbut covering a spectrum of the same affecting the thyroid gland.Thyroid is a trivial gland found at the base of neck, beneath Adam's apple. It yields two core hormones – Triiodothyronine: T3 and Thyroxine: T4 [1]. Thyroid, an important gland,controls how swiftly the body burns energy, makesproteins, and how subtle the body should be to other hormones.These hormones normalize the rate of metabolism and affect theprogression and rate of function of several other systems in the body.Too littleproduction of thyroid hormone causes Hypothyroidismnamely underactive thyroid; whereas too much production of the samehormone causes Hyperthyroidism called overactive thyroid [2].

Subsequently thyroid disorders present generic symptoms such as lethargy, increase in weight constipation etc., andare a lotchallenging to detect without testing. Due to lack of noticeable signs and symptoms of these disorders and lack of awareness, thyroid testing tends to be overlooked [3].

Over the past few years, there has been anenormousrise in the figure of thyroid cases.Diagnosis is an imperative task in medicinal science for of its criticality, efficacy and correctness in defining whether or not a sufferer has a specific disease. Data cleansing practices were smeared to create the data primeval enough for executing analytics to illustrate the threat of patients attaining thyroid [4].

LITERATURE REVIEW

Researchers in the latter years carried out a lot of work doneto diagnose the distinctsyndromes of thyroid by utilization of data mining techniques. These are theprocedures of machine learning including decision tree, random forest, SVM, naïve Bayes and ANN that are expansively utilized in the recurrentsyndromes and in the extrapolativeglitches.

In [5] methodical approach for prior diagnosis of Thyroid ailmentby means of back propagation algorithm in neural network is used. ANN determines in anuprightagreement with the primary data and specifies the advanced neural network which practices as a standby for prior syndromeforecasts.

In [6] researchers analysed and compared all four classification models specifically Naive Bayes, Multilayer Perceptron, Decision Tree and Radial Basis Function Network. The deduction reveals a significant precision for each of these models.

In paper [7], proposed a precise technique for detecting the thyroid by utilizing the back propagation algorithm. Artificial Neural Network is developed using the back propagation of error to identify the preliminary thyroid prediction. ANN is trained subsequently for testing the experimentally, but not the same training sets. The training can be done in two ways as supervised learning and unsupervised learning. The experimental result is carried out in MATLAB Neural Network Toolbox Software. This provides better performance than the simple gradient descent algorithm.

In paper [8], classification approaches are discussed that are utilized for the prediction of class label. This classification of dataset helpful for the prediction of various diseases from large volume of patient's dataset. Diabetic's dataset is used for the classification based on the decision table from the support and confidence to obtain better accuracy. The naïve bayes and fuzzy KNN are processed together for the medical dataset which provides better accuracy.

In paper [9], Ling Chen et al., proposed an algorithm called Semi-supervised Heterogeneous Graph on Health to predict the risky patients from the electronic health record. The Cause Of Death (COD) database are prepared the high risk dataset from the GHE database. The risky patients are classified using the semi-supervised learning with label propagation which includes the patient personal details, mental report and physical report.

Sudesh Kumar and Nancy [10] projected a clustering and data mining technique. The normalization approach is used to recover the efficient information with efficient factor. This approaches are Min-Max, Z-Scaling, decimal scaling normalization. The K-means clustering process helps to cluster in smaller time and to detect the similar clusters of objects centred on values of its attributes. The Z-score normalization and K-means clustering technique were integrated.

III. METHODOLOGY

In this work, 21 attributes of UCI thyroid dataset are conscripted. The dataset is being filtered by applying correlation based feature selection to reduce the 21 attributes to 14 attributes. The model is then built next to the selection of relevant features.

A) Feature Selection using High Correlation and Backward Elimination

A feature or an independent variable in any dataset may or may not have an effect on the dependent or output variable. The unrelated features may make the model worst. Hence features are sieved to select only the subset of relevant features using Feature Selection [11], a primary and imperative step while performing any machine learning task.

Correlation is a statistical word which in collective refers to in what way the two variables (A, B) are to closer and have a linear relationship among each other.

$$\text{Pearson's correlation coefficient} = \text{covariance}(A, B) / (\text{stdv}(A) * \text{stdv}(B))$$

Features with higher correlation are further linearly reliant and hence have more or less the same influence on the output variable. One of two features can be dropped when they are highly correlated.

The correlation coefficient values fall amid -1 to 1. A value nearer to 0 entails weaker correlation or simply no correlation, 1 implies robust positive correlation and -1 implies robust negative correlation [12]. The correlation coefficients amongst features are compared and removed one of two features that have a correlation higher than 0.9.

In the process of Backward Elimination, all the possible features are fed to the model at first. The conduct of the model is being checked to iteratively eliminate the worst performance features [13] one after one until the complete performance of the model arises in suitable range. The performance metric taken at this point to assess feature performance is p-value. If p-value of the feature is above significance level (0.05) then remove the feature, else we keep it.

B) Model Building

The model is built subsequently next to feature selection. The new dataset is split into two sets namely: train and test set. X% of the dataset is used as actual training set and left over (100-X)% as validation set. The model is iteratively trained and validated on these different sets. Classification models like Support Vector Classifier, k-

Nearest Neighbour and Decision Tree are used to build a model with K-fold Cross validation technique [14] that avoids over fitting.

(i) **Support Vector Machine:** The SVM handles multiple continuous and categorical variables. So, it is treated as both a classifier and regression method that maintains all the main features. The primary task of the algorithm is to forecast the class membership for categorical target by building hyper planes in a multidimensional space that splits cases of diverse class labels. SVM supports best prediction accuracy that eludes over fit. It also provisions text and sparse transactional information. The SVM offers empirically decent performance in the arena of bioinformatics, image and text recognition [15].

(ii) **k-Nearest Neighbour:** K Nearest Neighbors (KNN) is a modest algorithm which stocks altogether available cases to classify new cases centred on a similarity measure: distance functions. A case is categorised by a popular poll of its neighbours, with the case assigned to the class most common among its KNN, measured by a distance function [16]. An Euclidean distance among standard points: x and y is specified by the equation

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

The value of a positive integer, k is configured by scrutinizing the data set. It yields good result at $k=5$. This infers that the result will be further accurate when k value gets higher. In maximum cases, the optimal k value lies amongst 3 and 10.

(iii) **Decision Tree:**

Decision tree forms classification or regression model in a format of tree structure. The principle of splitting standards is behind the intellect of decision tree classifier. Decision trees are offered analogous to a flow chart, wherein instances are categorised rendering to their feature values. A node in a decision tree signifies an instance, outcomes of the test denoted by branch, and the leaf node epitomized the class label. It breaks down the dataset into minor subsets while incrementally developing an associated decision tree at the same time [17].

C) Performance Metrics of a Classifier

(i) **Accuracy:** - Accuracy is the utmost intuitive performance measure, a ratio of correctly predicted observation to the whole observations [18]. Accuracy is a great metric if there are symmetric datasets where values of false positive and false negatives are nearly analogous. For our model, an 0.9946 accuracy is got.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

(ii) **Precision:** - Precision is the ratio of correctly predicted positive instances to the total predicted positive observations. High precision relates to the low false positive rate. A precision of 0.9894 is achieved which appeals good.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

(iii) **Recall :** - Recall [19] or sensitivity is the proportion of correctly predicted positive observations to all the observations in actual class. A recall of 0.9893 has been obtained which is upright for this model as it is above 0.5.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

(iv) **F1 score:** - F1 Score is the weighted average of Precision and Recall. Therefore, this score receipts both false positives and false negatives into consideration. F1 is typically more expedient to accuracy, particularly if there is an irregular class distribution. Accuracy works superlative if false positives and false negatives have alike cost. If the cost of false positives and false negatives are very diverse, it is good to gaze at Precision and Recall together. In this instance, F1 score is 0.9893

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

(v) **Cohens Kappa:** - The Kappa statistic is a metric that relates an accuracy that is observed with the same that is expected. Computation of both accuracies is fundamental to conception of the kappa statistic, and is utmost effortlessly explained by using a confusion matrix [20].

$$\text{Kappa} = (\text{observed accuracy} - \text{expected accuracy}) / (1 - \text{expected accuracy})$$

D) Algorithm of the Proposed Model

Step 1: Load the dataset and normalize the dataset.

Step 2: Calculate the linear relationship coefficients between variables using Pearson correlation and drop one of the features that have high correlation (0.9)

Step 3: Select a significance level, say 5% (0.05)

Step 4: Consider the feature with the highest P-Value. If its P-value is greater than significance level ($P > SL$), discard the feature and go to step 4. Else the model is built.

Step 5: Split the data randomly into k -folds

Step 6: Fit the Decision Tree, SVC, k -NN Classifiers on ($k-1$) folds and validate them using k th fold.

Step 7: Measure the performance of the models for their accuracy.

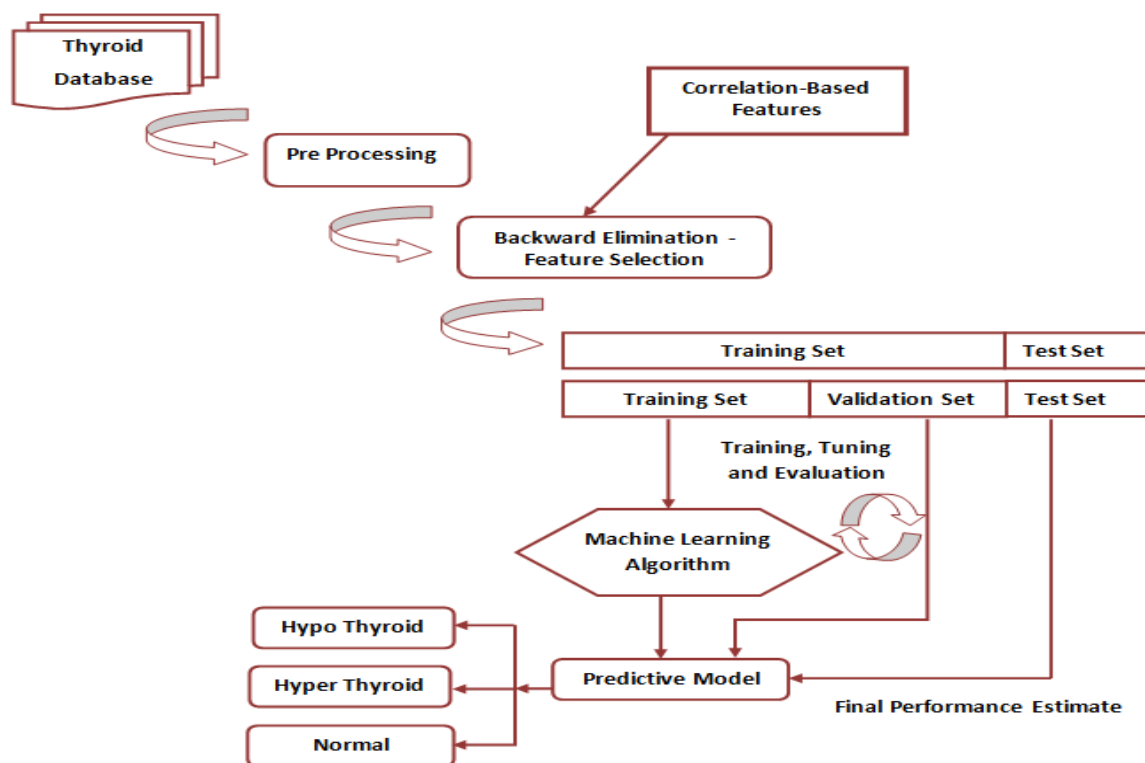


Figure 1: Workflow of the Proposed Model

IV. RESULTS AND DISCUSSIONS

To evaluate the performance of the anticipated methodology, experimentations are initiated using Python. The statistical analysis results obtained for correlation based classifications have been compared with each other and the ones taken for the whole datasets.

Data Description

UCI Thyroid dataset with 21 features and 7200 record instances of patients is taken into consideration. For the dataset, only factors concerning main characteristics were deliberated as initial attributes for feature scaling and their selection to diagnose the class of thyroid.

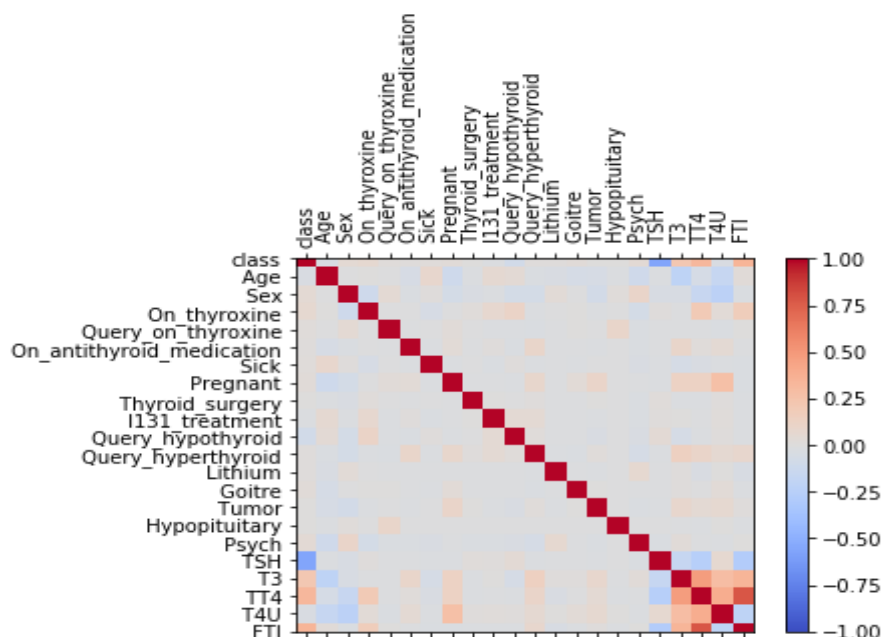
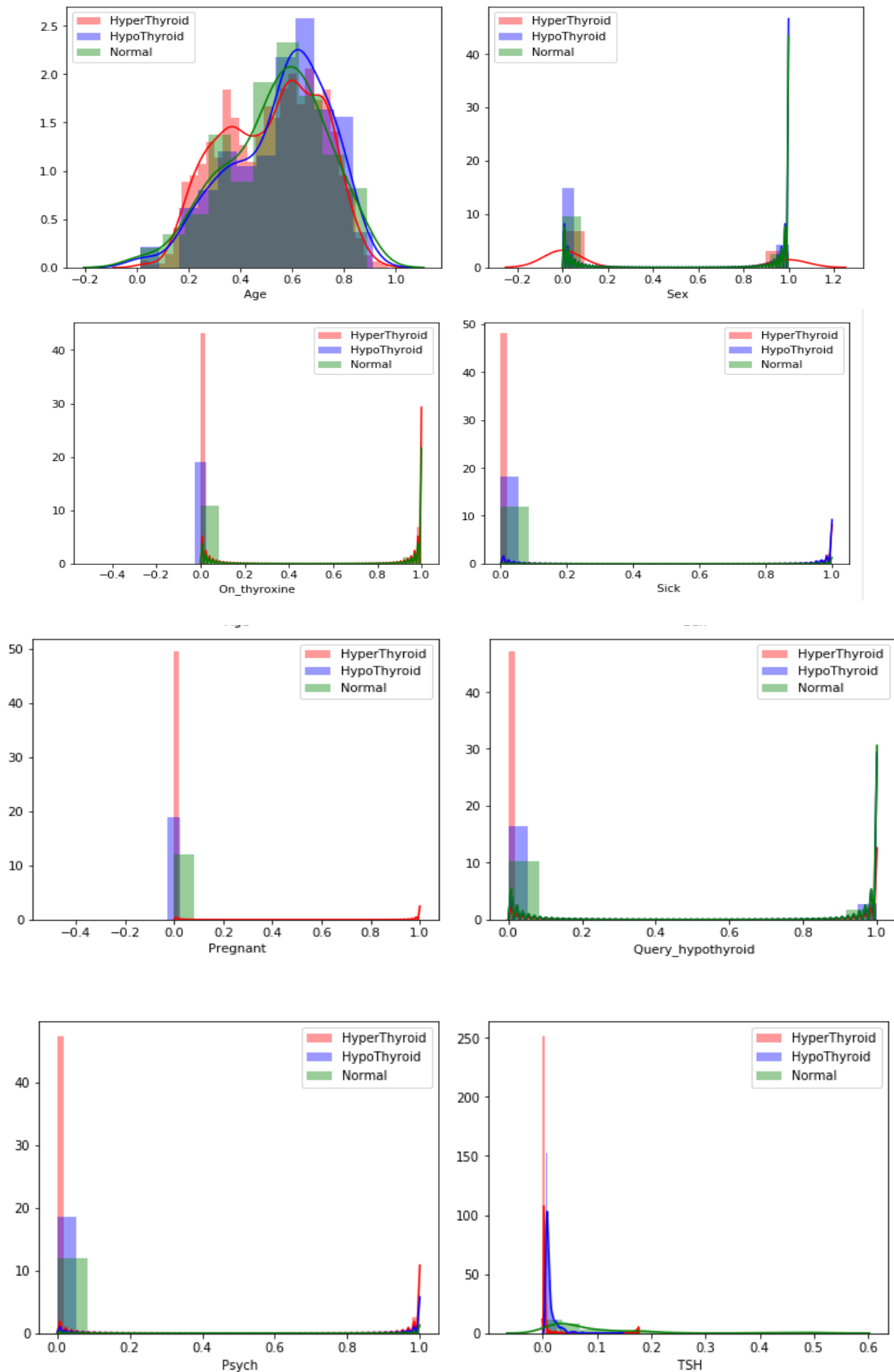


Figure 2: Correlation between every two independent variables

The correlation between features are compared to remove one of two features that have a correlation higher than 0.9. The features are then selected using backward elimination method based on p-value (0.05). Thus the dataset afterward feature selection comprises of 14 attributes with 7200 instances. The following figure shows the distribution plot of selected features.



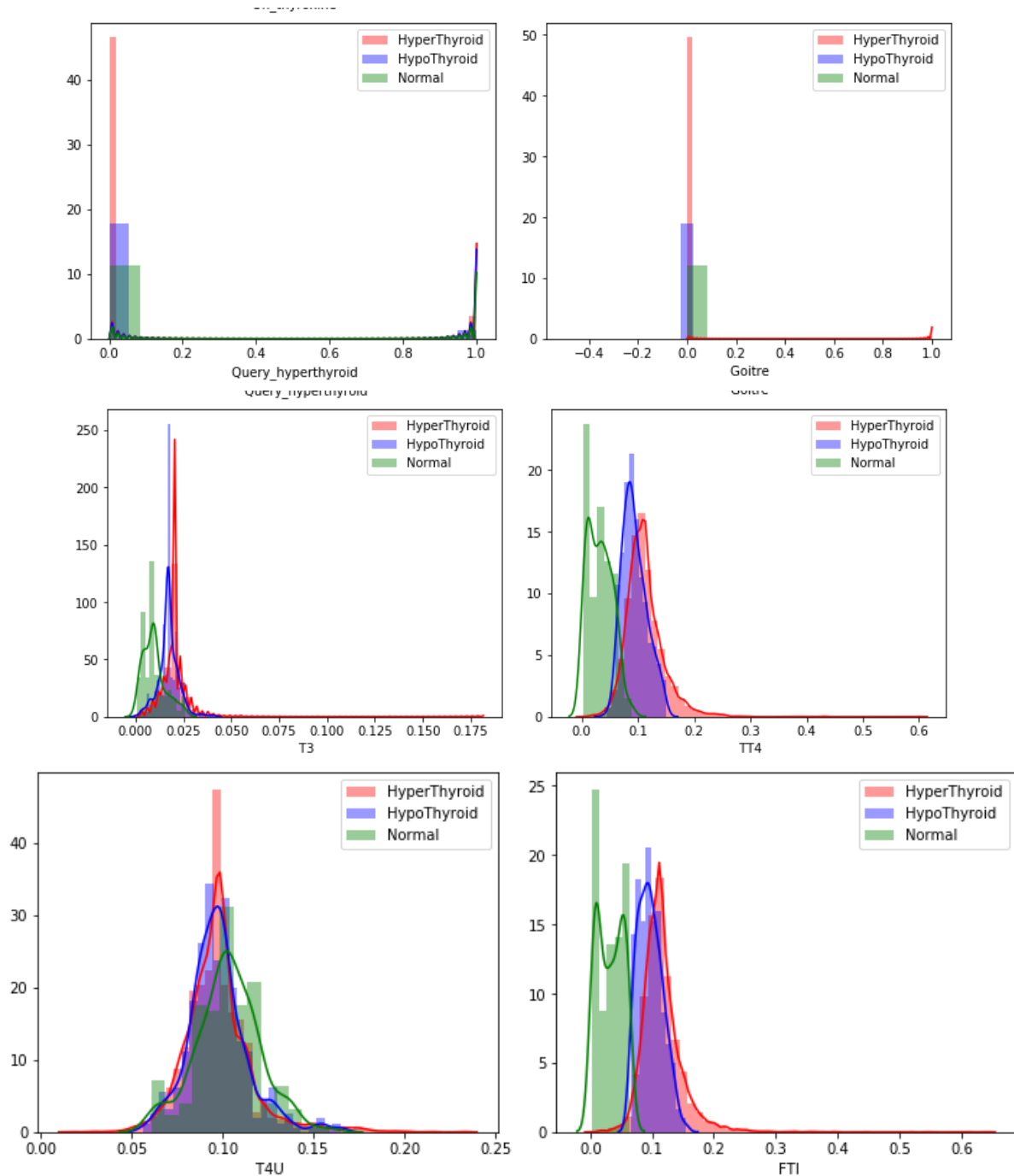


Figure 3: Visualization of selected features

Different classification models using Support Vector Classifier (SVC), k-Nearest Neighbour (k-NN), and Decision Tree(DT) were built by applying cross-validation technique to make accurate prediction of thyroid disease. It was proved that Decision Tree (DT) classifier gives an accuracy of 99.46% equalled to other classification algorithms.

Metric/Algorithm	k-NN	SVC	DT
Training Time (Sec)	0.031253	0.300804	0.015623
Accuracy	95.19	95.91	99.46
Precision	92.62	94.27	98.94
Recall	93.88	94.91	98.93
F1-score	92.28	93.91	98.93
Cohens Kappa	45.29	57.90	94.62

Figure 4: Performance Metrics for different classification algorithms

CONCLUSION

Prediction Models were constructed with and without feature selection. With the experimental results, it is observed that the accuracy of the predictions is enhanced when apt feature selection is implemented [21]. Envisaging the class of thyroid is probable with the help of machine learning by using diverse classifiers. Experimentally, the decision tree classifier was verified to be the fore-runners in terms of categorising the thyroid disease. From the distribution plots, it is obviously visible that the women are two times further likely to develop the risk of thyroid disease amid the ages between 18 to 42. The experimental result provides enhanced accuracy, precision, recall, F1-measure and Cohens Kappa evaluation with other classifiers: k-NN and SVM.

REFERENCES

- [1]. Yousif, N. M. Z. (2018). Estimation of Normal Thyroid Volume in Adults Using Ultrasonography (Doctoral dissertation, Sudan University of Science and Technology).
- [2]. Shroff, S., Pise, S., Chalekar, P., & Panicker, S. S. (2015, January). Thyroid disease diagnosis: A survey. In 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO) (pp. 1-6). IEEE.
- [3]. Sindoni, A., Rodolico, C., Pappalardo, M. A., Portaro, S., & Benvenga, S. (2016). Hypothyroid myopathy: a peculiar clinical presentation of thyroid failure. Review of the literature. *Reviews in Endocrine and Metabolic Disorders*, 17(4), 499-519.
- [4]. Tyagi, A., Mehra, R., & Saxena, A. (2018, December). Interactive Thyroid Disease Prediction System Using Machine Learning Technique. In 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 689-693). IEEE.
- [5]. S. SathyaPriya, Dr. D. Anitha" Survey on Thyroid Diagnosis using Data Mining Techniques" *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 6, Special Issue 1, January 2017.
- [6]. Kamble, M. S., Desai, A., & Vartak, M. P. (2014). Evaluation and Performance Analysis of Machine Learning Algorithms. *Neural Networks*, 2, 3.
- [7]. Ammulu, K., & Venugopal, T. (2017). Thyroid data prediction using data classification algorithm. *Int. J. Innov. Res. Sci. Technol*, 4(2), 208-212.
- [8]. Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC medical informatics and decision making*, 19(1), 211.
- [9]. Chen, L. (2017). Healthcare data mining from multi-source data.
- [10]. Kumar, S. (2014). Efficient K-Mean Clustering Algorithm for Large Datasets using Data Mining Standard Score Normalization. *Int. J. Recent Innov. Trends Comput. Commun.*, 2(10), 3161-3166.
- [11]. Mangal, A., & Holm, E. A. (2018). A comparative study of feature selection methods for stress hotspot classification in materials. *Integrating Materials and Manufacturing Innovation*, 7(3), 87-95.
- [12]. Jain, I., Jain, V. K., & Jain, R. (2018). Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Applied Soft Computing*, 62, 203-215.
- [13]. Li, H. D., Xu, Q. S., & Liang, Y. Z. (2018). libPLS: An integrated library for partial least squares regression and linear discriminant analysis. *Chemometrics and Intelligent Laboratory Systems*, 176, 34-43.
- [14]. Tripathi, M., & Taneja, A. (2019). K-Fold Cross-Validation Machine Learning Approach on Data Imbalance for Wireless Sensor Network.
- [15]. Abd, A. M., & Abd, S. M. (2017). Modelling the strength of lightweight foamed concrete using support vector machine (SVM). *Case studies in construction materials*, 6, 8-15.
- [16]. Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, 12(1), 90-108.
- [17]. Habrich, T., Wagner, C., & Hellingrath, B. (2018, July). Qualitative assessment of machine learning techniques in the context of fault diagnostics. In *International Conference on Business Information Systems* (pp. 359-370). Springer, Cham.
- [18]. Mallo, S. C., Valladares-Rodriguez, S., Facal, D., Lojo-Seoane, C., Fernández-Iglesias, M. J., & Pereiro, A. X. (2019). Neuropsychiatric symptoms as predictors of conversion from MCI to dementia: a machine learning approach. *International psychogeriatrics*, 1-12.
- [19]. Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D., & Lestantyo, P. (2019, October). Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data. In *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)* (pp. 14-18). IEEE.
- [20]. Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2017). In search of new product ideas: Identifying ideas in online communities by machine learning and text mining. *Creativity and Innovation Management*, 26(1), 17-30.

[21]. Awang, M. K., Makhtar, M., & Rahman, M. N. A. (2017). Improving Accuracy and Performance of Customer Churn Prediction Using Feature Reduction Algorithms. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 9(2-3), 127-130.